Nexus Linguarum

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

D3.1

Intermediate Activity Report. Working Group 3 “Support for Linguistic Data Science ”

Main authors: Dagmar Gromann, Thierry Declerck

1 November 2021

Project Acronym	NexusLinguarum
Project Title	European network for Web-centred linguistic data science

WG3 Title	Support for linguistic data science
COST Action	CA18029
Project Title	European network for Web-centred linguistic data science
Project Acronym	NexusLinguarum
Duration	48 months
Start of CA18029	October 2021
Project Website	https://nexuslinguarum.eu
Responsible Authors	Dagmar Gromann, Thierry Declerck
Contributors	Dagmar Gromann, Radovan Garabik, Ineke Schuurman, Thierry Declerck, Renato Rocha Souza
Reviewer	NexusLinguarum core group team
Version | Status	v1.0 | final
Date	1 November 2021

Acronyms List

CA	Cost Action
DL	Deep Learning
LD	Linked Data
LLD	Linguistic Linked Data
LLOD	Linguistic Linked Open Data
LOD	Linked Open Data
NLP	Natural Language Processing
MC	Management Committee
PRIMSA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
UC	Use Case
WG	Working Group

Table of Contents

1. Executive Summary

Working Group 3 (WG3) of the NexusLinguarum COST Action entitled ***Support for linguistic data science*** aims to foster the study of linguistic data by following data analytic techniques at a large scale in combination with LLD and linked data-aware NLP techniques. These techniques range from **Big Data** and **deep learning** to different linguistic description levels for **multilingual** and **multimodal** representations. Additionally, **education** in linguistic data science is one dedicated task of this WG.

This deliverable reports on the activities of WG3 as of M24 (October 2021) and focuses on meetings and events organized, surveys prepared, and publications as well as current and future plans for each individual task. In addition, collaborative and joint activities with other working groups and other initiatives, such as the OntoLex community are described. Finally, for each task an outlook for activities planned for the second half of this CA is presented.

2. Introduction

Support for linguistic data spans from investigating Big Data and deep learning to specific linguistic description levels for multilingual and multimodal modeling options of linguistic (linked) data. Furthermore, the activities of this WG provide support for education and educational programs for linguistic data science in a Web-centered context.

In NexusLinguarum linguistic data science is understood as a subfield of the rapidly growing field of data science. Data science can be described as the systematic analysis and study of the structure and properties of data, including methods and techniques to extract knowledge and gain insights from data. The subfield of linguistic data science investigates the analysis, representation, integration, and exploitation of linguistic data for language analysis and language applications. Language analysis spans different linguistic description levels and theoretical bases, e.g. syntax, morphology, terminology, lexicology, etc. Language applications relate to common NLP tasks, e.g. machine translation, speech recognition, sentiment analysis, etc. Linguistic data are typically contained and described in language resources.

Within this context, WG3 aims to cover the broader topic of support for linguistic data science within NexusLinguarum and this document presents first the objectives of WG3, its structure and task leader profiles and the main preliminary outcomes and publications that resulted from WG3 activities over the period of the first two years of NexusLinguarum from October 2019 to October 2021. The document then continues to detail the activities for each individual task prior to the brief conclusion of this document.

2.1. WG3 Objectives

The title and task of WG3 is to provide support for linguistic data science. The main objective is to provide support in the form of information of existing and needed resources, approaches and standards. In more detail, the following major objectives have been formulated:

- collect resources and approaches, especially regarding
  - Big Data and LLD
  - Deep Learning and LLD
  - multilingual modeling and LLD
  - multimodal and multidimensional modeling and LLD
- prepare comprehensive state-of-the-art reports on specific topics
- conduct surveys on the utilization and support of deep learning for LLD
- propose, collect, publish and report on modeling of multilingual and multimodal aspects in linguistic data science

- collect and report on skills and competencies important for and in linguistic data science
- propose training programs for linguistic data science

2.2. WG3 Structure and Core Group Profiles

WG3 is one of four WGs within NexusLinguarum with approx. 68 registered members, six task leaders and five tasks, of which one with two subtasks. The profiles of WG3 participants range from linguists to computer scientists, including deep learning specialists. An overview of the variety of profiles is reflected in the core group profiles provided below. An overview of the structure of WG3 is provided in Fig. 1.

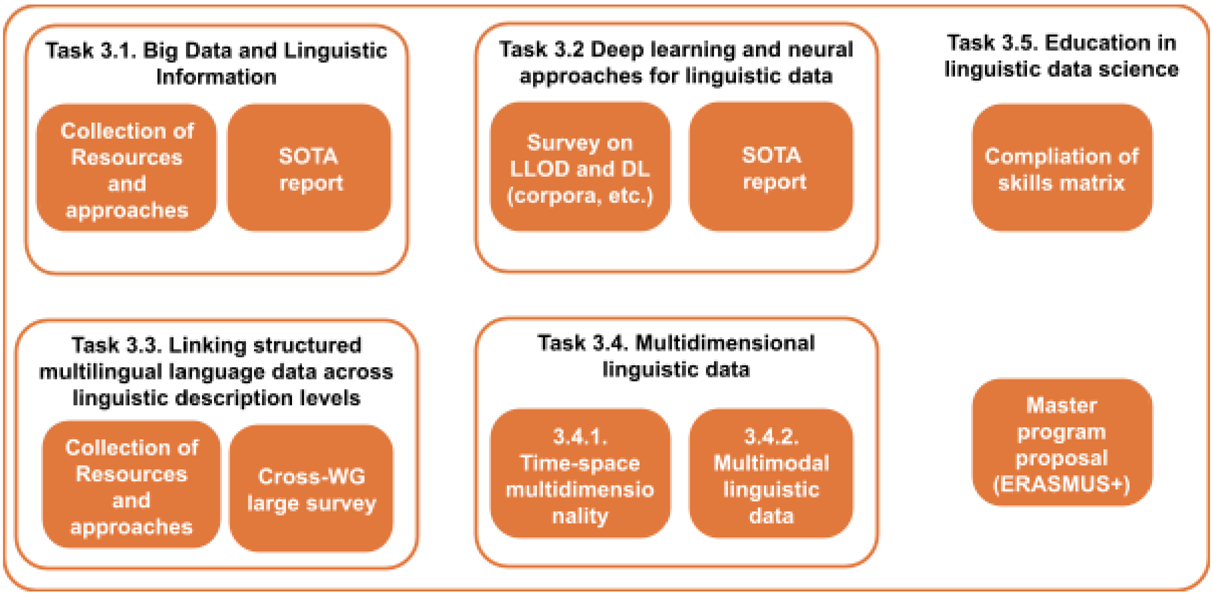

Task 3.1. Big Data and Linguistic Information
Collection of Resources and approaches
SOTA report
Task 3.2 Deep learning and neural approaches for linguistic data
Survey on LLOD and DL (corpora, etc.)
SOTA report
Task 3.5. Education in linguistic data science
Compliation of skills matrix
Task 3.3. Linking structured multilingual language data across linguistic description levels
Collection of Resources and approaches
Cross-WG large survey
Task 3.4. Multidimensional linguistic data
3.4.1. Time-space multidimensio nality
3.4.2. Multimodal linguistic data
Master program proposal (ERASMUS+)

Figure 1: Structure of WG3 and overview of all tasks

In the following a detailed account of the WG3 core group profiles, that is, WG leader, WG co-leader, and task leaders, is provided.

Dagmar Gromann, University of Vienna, Austria (WG3 leader and Task 3.3. leader): is Assistant Professor Tenure Track at the Centre for Translation Studies of the University of Vienna with a focus on computational terminology and language technology. Her research particularly focuses on machine learning and deep learning approaches to multilingual information extraction, including terminological concept systems and cognitive linguistic concepts. She has been project leader of the pilot project “Extracting Terminological Concept Systems from Natural Language Text” (Text2TCS) funded by the European Language Grid (ELG), available as an ELG service, currently leads the project on gender-fair machine translation (GenderFairMT) and is a member of the European Language Equality (ELE) project. She is National Anchor Point for the European Language Resource Coordination (ELRC), National Competence Center (NCC) for ELG and has years of experience in LLD and recently has started working on deep learning and LLD. She has organized multiple international scientific events, including the 3rd Conference on Language, Data and Knowledge (LDK) 2021 in Zaragoza, Spain, supported by NexusLinguarum in the role of Program Committee Chair.

Thierry Declerck, DFKI, Germany (Science Communication Manager, WG3 co-leader and Task 3.4.2. co-leader): is senior consultant at the Multilinguality and Language Technology (MLT) Lab of the German Research Center for Artificial Intelligence (DFKI GmbH) in Saarbrücken, Germany. He has been working in a series of European and national projects dealing with a broad range of NLP topics. He is currently in charge of the DFKI contribution to the H2020 Prêt-à-LLOD Project, which is dealing with the topic of "Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors". Thierry is also instrumental to two W3C Community Groups, Ontology Lexica (Ontolex) and Linked Data for Language Technologies (LD4LT). Together with Dagmar Gromann (and other colleagues), Thierry has co-organized a series of workshops on "Semantic Deep Learning". Thierry was co-chair of the LDK 2021 conference.

Radovan Garabik, Slovak Academy of Sciences, Slovakia (Task 3.2. leader): works in the field of corpus linguistics, natural language processing and digital lexicography. He was responsible for the design and implementation of written Slovak corpus, corpus of spoken Slovak, Slovak morphology analyser, parallel corpora, Slovak Terminology Database, on-line Slovak language dictionaries portal, and a database of Slovak language linguistic resources. He is also participating on corpus of syntactically annotated Slovak language, corpus of Slovak dialects, several monolingual and bilingual dictionary projects, machine translation projects and is the principal editor and author of several specialized dictionaries and language processing tools and resources. Currently, he is a member of the ELE project, a contact person for the National Competence Center NCC for ELG and represents Slovakia in COST Action CA19102 'Language in the Human Machine Era' and EU – CEF action Curated Multilingual Language Resources for CEF.AT.

Ineke Schuurman, KU Leuven, Belgium (Task 3.4.2 leader): as a voluntary research associate she is member of the Centre for Computational Linguistics (KU Leuven) where she started in 1989 as delegate coordinator of Eurotra-Leuven (MT). Thereafter she was involved in many national, international (Netherlands - Belgium), and European projects. From the beginning (2008) till her retirement she was also involved in CLARIN (ERIC) in several positions. Lately she co-ordinated the EU Competitiveness and Innovation Framework project "Able to

Include". Currently she is also involved in the COST Action "advancing Social inclusion through Technology and EmPowerment" (a-STEP) and in the EU project "Sign Language Translation Mobile Application and Open Communications Framework" (SignON).

Renato Rocha Souza, Austrian Academy of Sciences, Austria (Task 3.5. leader): works in the Austrian Academy of Sciences - Austrian Centre for Digital Humanities and Cultural Heritage; in the Department of Image Science, in the Donau Universität at Krems; as a lecturer in the Universität Wien and as a professor and Researcher at the Fundação Getulio Vargas, Brazil. He has been a Faculty Professor and Researcher for 21 years in the Computer Science, Data Science, Applied Mathematics and Information Science domains. Has been participating as a PI and as a member on research projects for the past 10 years in the domains of Data Science, Knowledge Organization, Knowledge Management, Open Innovation, Political Science, Law, Economics, Smart Cities, Public Health and Education. Was co-director of the "Archives Without Borders" Project, headed by Columbia University. Was also part of the Board of the International Society for Knowledge Organization (ISKO) for the past 8 years.

2.3. Preliminary Outcomes and Publications

In order to describe our activities in the form of measurable outcomes, this section details events organized within and for WG3, resources collected and provided by WG3, and publications that have resulted from our activities.

Events:

In terms of **meetings**, we organize regular WG3 video conferences every two months to welcome new members that keep on subscribing and to discuss updates from the individual tasks and from NexusLinguarum as a project. It is also the venue for us to discuss further cross-WG collaborations and invite other initiatives to join our activities. Additionally, each task has been organizing regular online videoconferences (mostly once a month) to discuss specific details related to a subtopic of the specific task. In addition, we have been organizing joint WG telcos, especially with WG1 and WG4 leaders and members, to further existing collaborations. The yearly Management Committee meeting allows us to share our activities and findings with the whole COST Action and further discuss joint WG activities with all NexusLinguarum members.

For most topics covered within WG3, a collection of resources and a survey on the state-of-the-art of approaches and publications was most reasonable. However, for the topic of deep learning and linguistic (linked) data we opted for alternative methods due to two main reasons: (1) the field is extremely fast-lived and a lengthy scientific survey method might be outdated by the time we complete it, (2) NexusLinguarum as a whole represents the linguistic linked data community very well and we first wanted to obtain a full picture of deep learning approaches and activities within the field. We, thus, decided to start activities in Task 3.2. with a **survey** focusing on the utilization of deep learning and linked data approaches by central European data providers. This survey was also distributed beyond the NexusLinguarum community. To obtain a good overview of research conducted on this topic within the NexusLinguarum community, we organized the First **Workshop** on Deep Learning and Neural Approaches for Linguistic Data collocated with our Third MC Meeting in Skopje.

Resources:
In terms of resources, we shared the Task 3.2 **survey results** on deep learning, linked data, and linguistic data with the entire COST Action and also provided a book of abstracts (Radovan 2021) for the **Workshop** on Deep Learning and Neural Approaches for Linguistic Data that is publicly available.

At the very beginning of NexusLinguarum, we collected a repository of resources for all subtopics of WG3 shared with all members. This repository represented a **collection of existing tools, resources, and standards** for Big Data, Deep Learning in connection with linguistic data, linguistic (linked) data in general, and specific to multilingual and multimodal representations and models. For Task 3.5. we initially compiled a separate repository of all educational initiatives - from Massive Open Online Courses (MOOCs) and tutorials to bachelor and master programs - related in any way to linguistic data science. After having compiled an initial list within WG3, we then called on all NexusLinguarum members to add further initiatives we had missed. Based on this list of initiatives, we compiled a list of competencies and skills central to **linguistic data science** in the form of a **matrix of competencies** grouped by topics and where, how, and how often these are covered by educational activities. We then ranked the compiled competencies and skills by importance to our endeavor of preparing a joint master program on linguistic data science, the next step for this activity.

Furthermore, within the work on multimodal modeling (Task 3.4.2) a first **ontology on data categories for sign languages** has been created as a first step to bring benefits of interoperability and facilitated sharing offered by linked data to sign languages. This is ongoing work and the ontology is currently still being refined but will soon be publicly available.

Publications:
The first publication for the Task 3.2. is the Book of Abstracts from the **Workshop** on Deep Learning and Neural Approaches for Linguistic Data (Garabik 2021), which is available here

and contains all abstracts submitted to the workshop. In this Book of Abstracts, the contributions are multilingual term extraction based on neural language models (Gromann et al. 2021), speaker attitude detection by analysing discourse markers with a neural language model (Oleškevičienė et al. 2021), extracting relational knowledge from masked language models trained on Portuguese (Oliveira 2021), neural named entity recognition on Romanian legal language (Păiș & Mitrofan 2021), and using neural object detection methods for analysing the document structure of academic publications (Susman et al. 2021). This range of tasks and approaches relating linguistic data science with deep learning shows well the multi- and interdisciplinary character of this working group. Apart from this workshop, members of WG3 in general investigate this relation, e.g. Rokas et al. (2020) investigate several neural architectures and neural language models to perform automated term extraction in the cybersecurity domain in Lithuanian.

Another major publication that is to be submitted by the end of this calendar year is a substantial and **systematic state-of-the-art review on multilingual linguistic linked data**. In total 19 experts jointly collaborated to prepare this publication to be submitted to the Semantic Web journal. Following the PRISMA guidelines for reviews we have performed a systematic search and review of existing approaches, detailed below in the Task 3.3. activities.

Furthermore, we intend to prepare a **scoping review on Big Data and linguistic data science**, which has undergone the first search stage and is now in the process of reaching the paper identification and summarization/review stage. The publication venue for this activity has not yet been identified.

In terms of **multimodal linguistic data science**, Declerck and Bajčetić (2021) published a paper for adding pronunciation information derived from the English edition of Wiktionary to the Open English WordNet. On a more specific topic, Declerck (2020) published a paper on extending the Open Dutch WordNet with Dutch lexicographic resources.

A published collection of interviews by the Terminology Coordination Unit of the European Parliament (2021) contains an interview with Dagmar Gromann by Justyna Dlociok. It is an **interview on terminology**, where Dagmar refers to linked data science and briefly explains what we want to achieve in regards to terminology within NexusLinguarum.

One general publication on **recent developments of the LLOD infrastructure** by a number of NexusLinguarum members and external colleagues was published at LREC in 2020 (Declerck et al. 2020). NexusLinguarum members and OntoLex collaborators also organized a

Workshop on Linked Data in Linguistics at LREC and published its proceedings (Iono et al. 2020).

In collaboration with WG4, a number of papers have been submitted for publication on the use case of cybersecurity (Task 4.3.1), including Utka et al. (forthcoming) on bilingual term extraction in the cybersecurity domain and Rackevičienė et al. (submitted) on a methodology for building an English-Lithuanian corpus for term extraction and Rackevičienė et al. (submitted) on a detailed analysis of term distribution across genres. Another cross-WG collaboration by Armaselu et al. (2021) resulted in an LDK publication on the history of socio-cultural transformations as LLOD.

2.4. Collaborative and Joint Activities with other Groups and Initiatives

WG1 & WG4:

Collaborations with WG1 and WG4 currently mainly take place as explicit and intensive exchange between task leaders and members for multilingual modeling (Task 3.3. and Task 1.3/Task 1.1.). Furthermore, Task 1.1. and Task 3.4.1 collaboration has been initiated on modeling aspects of time and space in linguistic linked data, an activity that also involves UC4.2.1 and UC4.2.2. members and activities on diachronic modeling. For the topic of Big Data, mainly members of WG4 and WG1 with experience on the topic have joined the initiative. On the use case of cybersecurity there is collaboration between members from Task 3.2 and Task 4.3.1 for publishing joint work.

WG2

New collaborations on use cases for deep learning and linguistic data science are foreseen between Task 3.2 and two to three tasks of WG2 (especially on knowledge extraction and machine translation). This work has only been initiated after the last MC meeting.

OntoLex-Frac

WG3 closely collaborates with members of the W3C Community Group OntoLex, especially with members of the interest group on OntoLex module for FRequency, Attestations and Corpus data (FRAC) on the topic of multimodal modeling of linguistic linked data (Task 3.4.2 of WG3). Furthermore, additional collaborations with OntoLex members on modeling time and space in reference linguistic linked data have been initiated.

The Multi3Generation COST Action CA18231

Thierry Declerck was invited on the 6th of October 2021 to give a talk entitled “About the Linguistic Linked Open Data initiative” at the plenary meeting of the Multi3Generation COST

Action CA18231 (https://multi3generation.eu/). The purpose of this invitation was to investigate potential lines of cooperation between Multi3Generation and NexusLinguarum. Two topics for possible cooperation were immediately recognized: the encoding of Sign Languages and the use of knowledge graphs for the generation of common sense inferences.

LD4LT meetings

NexusLinguarum members are involved in and leading a W3C Community Group: The Linked Data for Language Technologies (LD4LT) that was founded in the previous FP7 project “LIDER'' and that is used for the broader discussion of issues related to linked data and its applications in NLP. NexusLinguarum partners are participating to the regular LD4LT telcos and a LD4LT annotation workshop, as a satellite event to the 3rd Language, Data and Knowledge (LDK 2021) conference, has been recently co-organized by those NexusLinguarum members.

3. WG3 Task Activities

This section provides details on all WG3 activities structured by task, providing information on the task leader, a brief overview of the task, and then detailed explanations of all activities.

3.1. Task 3.1. Big Data and Linguistic Information

Task leader: Konstantinos Tsagarakis recently stepped down, new task leader to be appointed soon

Overview:
In this task, Big Data sources and state-of-the-art statistical analysis are studied in combination with LLOD in order to better understand language. Visual analytics will be also considered for this task. This will have an impact on all subdomains of linguistics, from typology to syntax to comparative linguistics.

Activities:
The first activity of Task 3.1 was to first discuss the term ``Big Data’’ within the context of NexusLinguarum, which we specify as a term that particularly relates to the size and complexity of the dataset. Size refers to the fact that the data are difficult to process with standard approaches (also one V, the volume) and complexity has been denoted with 5 to 17 Vs, such as variety, velocity, veracity, visualization, within the context of Big Data research. One major point of discussion at this point was the differentiation of static and dynamic data. Static would be equivalent to a corpus or other language resources that are once compiled and then published. For Big Data, the phenomenon of dynamic data exists, i.e., data which are continuously updated and extended from various sources.

As a second and main activity we decided to conduct a survey on state-of-the-art approaches combining Big Data and LLD. To this end, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, we first specified a list of search terms to query scientific databases, e.g. [“Linked data” AND “semantic” AND “big data”], and the search platforms, which are:

- Scopus search
- Web of Science search
- Google scholar
- DBLP
- Research gate; Academia
- standards: W3C, ISO, SAE
- conference proceedings: LDK, LREC, ACL, EMNLP, COLING, EACL, IJCAI, ECA

For instance, the above keyword resulted in a total of 77 publications on these platforms. Sorting the returned publications by authors and by relevance of contribution, it became rapidly evident that the topic of LLD and Big Data is relatively novel with a limited number of relevant publications, whose main contributors are members of the NexusLinguarum network. Thus, we are now in the process of turning the initial idea of a systematic review towards a scoping review that aims to identify the scope and nature of the existing literature. After a break due to changes in task leadership, we will continue this task with regular telcos to finish the already started collection and scoping of existing approaches, such as Janev et al. (2020), where the first author is also a member of NexusLinguarm.

3.2. Task 3.2. Deep learning and neural approaches for linguistic data

Task leader: Radovan Garabik

Overview:
Currently, deep learning techniques have gained popularity in many research areas, NLP being one of them. In fact, the field is being revolutionized by the emergence of relatively available huge language models based on attention (transformers), such as BERT and variations (and less available GPT-x models). The goal of this task is to study the effective use of deep learning in understanding the specifics of linguistic data in a big data context, to be better exploited and combined with linked data mechanisms.

There are two separate approaches involved in the task:

- Using LLOD as a platform to exploit or populate in NLP
  - Publishing and accessing deep learning models, results, tools as linked (open) data
  - Using linked data to train deep learning models
- Using deep learning on the structure of LLOD interlinking itself. This is currently only in the nascent phase.

Task 3.2 is coordinated by regular monthly online meetings.

Activities:

3.2.1. Survey

We carried out a survey on methods of automatic corpora language analysis, aimed to map currently used methods for fundamental NLP text analysis as used in the context of major large text corpora. The survey distinguished between rule-based, statistical and deep learning methods respectively and the invited participants were major providers of large corpora within Europe.

While not generally representative for European languages, the survey results show that deep learning methods are gaining traction and are used more and more in all areas of NLP as used in corpus linguistics. On the other hand, “basic” NLP tasks, such as tokenization, sentence segmentation, lemmatization, part of speech tagging and morphological description are still dominated by rule based or statistical methods and people are in general satisfied with the accuracy.

Below is a selection of survey results, an overview of languages of the corpora included in the survey (Fig. 2); satisfaction of the participants with Deep Learning methods used for various NLP related tasks (Fig. 3);satisfaction of the participants with the use of Linked Data for various linguistic tasks (Fig. 4).

Detailed results are in:
https://docs.google.com/document/d/1r1ZImSlDYxe8wBqY1bFxuva9AoNJ2ds03ZdPPi08LKg/

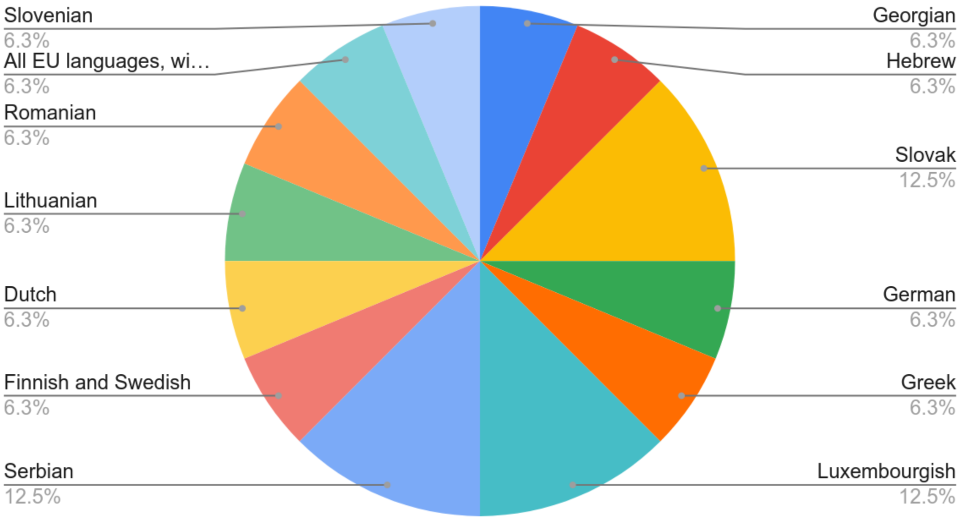


Figure 2: Languages of the corpora included in the survey

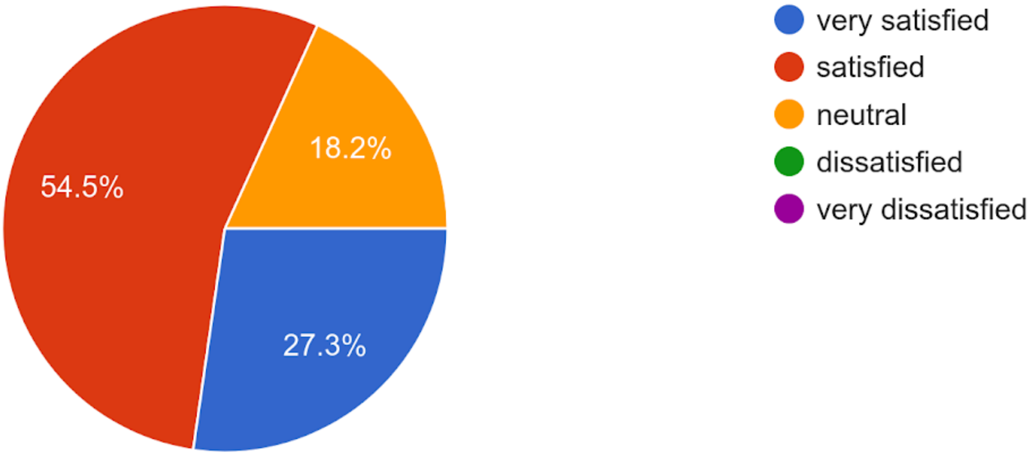


Figure 3: Satisfaction with Deep Learning methods for processing linguistic data

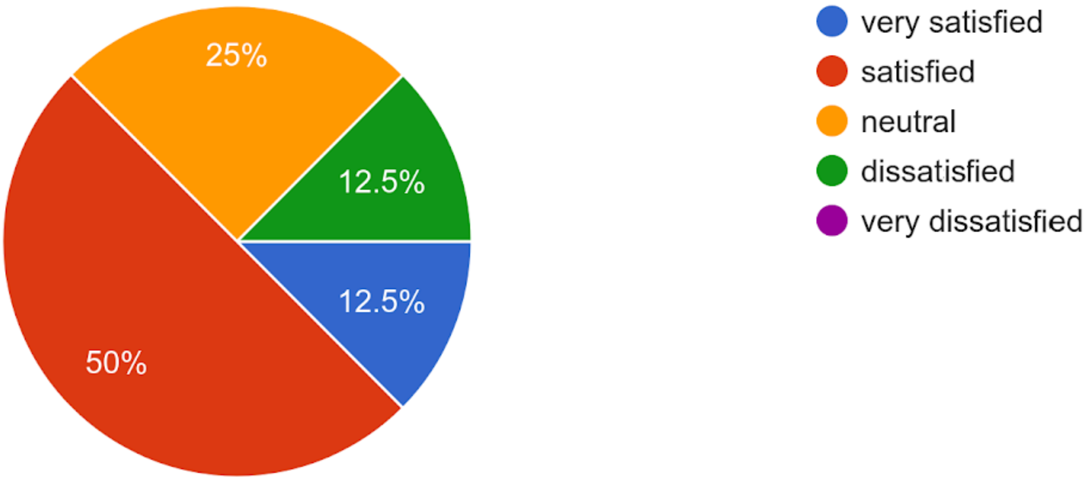


Figure 4: Satisfaction with the use of Linguistic Linked Data

3.2.2. Workshop

We organized a *Workshop on Deep Learning and Neural Approaches for Linguistic Data* held in hybrid mode in Skopje, North Macedonia; and online via the Zoom conferencing platform on 30 September 2021.

The workshop was aimed at deep learning in connection with linguistic data and the effective use of deep learning in understanding the specificities of linguistic data. The submissions concerned deep learning used to improve named entity recognition; BERT in conjunction with a compilation of lexical patterns to automatically acquire lexico-semantic relations; using transformer models to predict discourse relations and speaker's attitudes; using transformer models to automatically extract terminological concept systems; and an automatic detection of rhetorical patterns in academic texts using machine learning algorithms designed for image object detection purposes trained on the page layout and graphical elements.

The workshop covered a fraction of the variety of problems that modern deep learning methods can successfully address, and demonstrated the usefulness of linguistic linked open data, as results of and interconnected with neural approaches.

Workshop webpage and more details:
https://www.juls.savba.sk/workshop_20210930_en.html

Proceedings of the workshop are published in online form in a book of extended abstracts (Garabík 2021), available at:
https://www.juls.savba.sk/attachments/workshop_20210930_en/workshop_20210930_book_of_abstracts.pdf
and on Zenodo NexusLinguarum community:
https://zenodo.org/record/5566237

The panel discussion on the topic of Deep Learning and Linked Data touched several current issues, and these are summarized in the following text.

There seem to be certain issues in Deep Learning that a multilingual LLOD could help with, because large models (such as GPT-3) are difficult to train, require a vast amount of data, and computing infrastructures, which frequently are only affordable by large corporations. One way of reducing the amount of training data required and the computing time, which also reduces our carbon footprint, is to utilize structured resources in the fine-tuning and training process, which is where data available in the LLOD cloudtype – further specifies the type of the document; here: ORIGINAL_SCIENTIFIC_TEXT or REVIEW. could be useful. For such a scenario, a use case has been discussed: successful transfer learning from one language to another using existing LLOD data (dictionary from the Apertium project), as such a re-use reduces the need to re-train or train models from scratch. There is a great opportunity to look for synergies between Deep Learning and LLOD - we can combine DL and LLOD, use DL to create LLOD; coming from the other side, we could integrate information coming from DL to be used in larger knowledge based systems.

Many participants (of NexusLinguarum) started with Semantic Web and Linked Data and then turned their research attention to machine learning methods and/or NLP. Thus, several researchers, and also NexusLinguarum members, are proficient in LLOD and DL, which represents an excellent opportunity to strengthen the investigation of mutually beneficial synergies between both fields. The recent success of DL models trained on large datasets has provided astonishing results in most NLP tasks, however, for new tasks it is mostly still necessary to find a way, automated or manually, to annotate data.

One way of joining together DL and LLOD is the opportunity to use Linked Data as input to DL models: given recent advancements in graph neural networks we can conceive of applying DL to knowledge graphs, get embeddings on nodes, etc. Probably the distance between these two worlds will become smaller and smaller with the advances in the field. DL methods offer a creative way of playing with language data and it is worthwhile to explore the relations between DL and LLOD.

The discussion touched upon the challenge of bias and debiasing language models. Huge models propagate the bias in their initial training after they are fine-tuned. For instance, a pretrained model fine-tuned on relation extraction predicts no relation between the words *white* and *criminal* and synonymy for *black* and *criminal*. High-quality and large sets LLOD relations might help the debiasing process of such models.

What we see in databases is not what we typically assume in real life discussion, we very often do not mention things that are commonly understood; the base of what is commonly understood cannot be derived just from looking at the data. One of the often repeated issues in the NLP community is the recent advancement of data science (in the general sense) and its application to language. In computational linguistics, it has become frequent to optimize existing models. Even with little experience in DL or little to no background in linguistics, it is possible to re-use pretrained models and with such contributions we might only gain insights into model behavior or learned representations rather than deep research insights into linguistics or the task at hand. And it appears some of the tasks where DL has very good results are biased towards their susceptibility towards DL. Since fine-tuning pre-trained models frequently leads to state-of-the-art results in many NLP tasks, there is a danger of reducing the variety of research in (computational) linguistics, which, however, might only be a current trend. In any case, the consensus of the panelists was that LLOD can contribute by offering additional knowledge to DL models, including commonsense reasoning and natural language inferencing.

3.3. Task 3.3. Linking structured multilingual language data across linguistic description levels

Task leader: Dagmar Gromann

Overview:
This task focuses on how data for the basic levels of phonology, morphology and lexicon, often spread across datasets of varying extent, quality and format, can be described, stored and accessed uniformly.

Activities:
Starting from the idea of providing a best practice for modeling different linguistic description levels in relation to LLD, we quickly realized that a state-of-the-art review is required in order to evaluate work that has already been done on the topic and then work towards a best practice on that solid basis. Thus, we decided to apply the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to this task of providing a systematic review similar to Task 3.1.

In total, 19 experts are involved in the preparation of this systematic review, who jointly identified 41 important keywords for the search of relevant publications, e.g. [“multilingual LLOD”]. A total number of six experts then ranked the keywords on a scale from one to ten. The average score of that ranking then provided an overall score for each keyword. We searched across four main scientific search platforms, which are Scopus, Web of Science, DBLP, and Google Scholar within the time period of 2019 to 2021. A total of 25,074 publications resulted from this search. To manage this massive result corpus, we applied the keyword ranking in combination with the number of occurrences of publications across search platforms. We then annotated the top-most 210 papers regarding their relevance as well as potential subtopic of this topic. Thereby, we could identify the best ranked and humanly determined top 112 papers that were then clustered by subtopic in order to start a detailed reading and systematic review. To further validate this automated search, we also compiled a repository of publications that these experts considered central to this topic that was compared against the automatically generated results with a very decent overlap, that is, 80% of the publications considered relevant by human experts could also be retrieved by our automated search and semi-automated ranking method.

We decided to determine the number of experts needed for reviewing and summarizing the identified subtopics based on the number of papers that resulted in each, which are represented in Table 1. For the largest topic, OntoLex, three experts were assigned, while for smaller subtopics one to two experts reviewed the resulting papers.

Subtopic designation	No Papers	No Experts
application	15	2
BabelNet	5	1
Literature review	5	1
LLOD infrastructure	4	1
morphology	5	1
ontolex-lemon	25	3
overview paper	6	1
representation	12	2
resources	12	2
standards	5	1
under resourced languages	4	1
use cases	12	2

Table 1: Types of subtopics and number of papers of the result set as well as experts assigned to each subtopic

We have finalized the summarization of publications relevant to each subtopic, including an additional human search for publications the automated method might have missed. To this end, each topic was explicitly assigned to experts experienced in the relevant subtopic. We have identified the main linguistic description levels to this date covered by existing approaches, which are represented in Figure 5 where the size of the bubble approximates its coverage in existing publications.

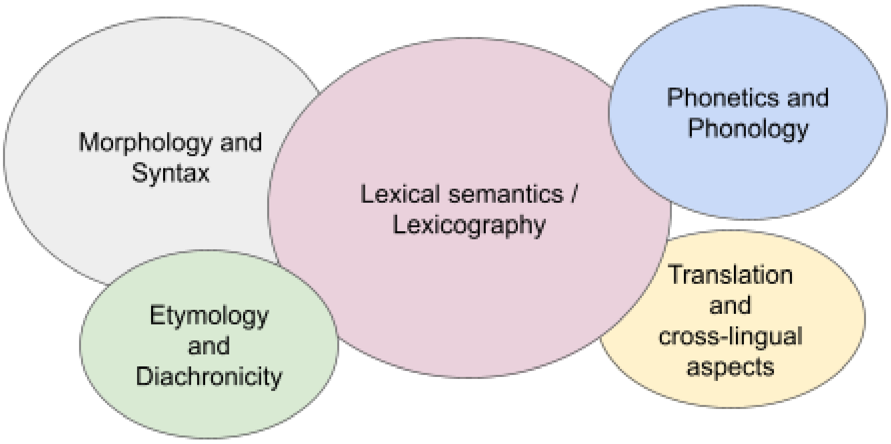

Morphology and Syntax
Lexical semantics / Lexicography
Phonetics and Phonology
Translation and cross-lingual aspects
Etymology and Diachronicity

Figure 5: Representation of linguistic description levels in final result set

We have additionally documented and identified the major challenges for multilingual linguistic linked data and Web-centered linguistic data science, which range from challenges regarding tool support to data protection and legal issues, and have compiled an expert-based guide for an ideal ecosystem for multilingual LLD. The submission of this systematic review paper is foreseen for the end of 2021 the latest.

3.4. Task 3.4. Multidimensional linguistic data

The original idea of this task as proposed initially was to focus on various dimensions of linking language resources, e.g. time and sociolect, in order to foster diachronic and sociolinguistic interoperable research across different sources in LLOD. However, it turned out to be practically very difficult to define these multiple dimensions within one task and some dimensions, especially diachronic aspects, strongly overlapped with use cases in WG4. Thus, we finally decided to split Task 3.4. into two tasks that foster the clarity of activities and reduce overlap with other activities in NexusLinguarum. The first subtask will provide a container task to join activities of WG1 and WG4 on time and space. The second subtask focuses on multimodal modeling of LLD.

3.4.1. Task 3.4.1. Time-space multidimensional linguistic data

Task leader: To be appointed

Overview:
The objective of this task is to bring members and activities of WG1, WG4, and WG3 on modeling aspects of time and space in linguistic (linked) data together in one overarching task. Due to our focus to first initiate aspects of multimodality, this subtask has only recently been initiated.

Activities:

A first collection of members of WG1, WG3, and WG4 interested in working on this topic has been evaluated in the joint MC meeting sessions across working groups. To properly initiate a discussion of this topic, we will organize our first larger video conference in November.

3.4.2. Task 3.4.2. Multimodal linguistic data

Task leaders: Ineke Schuurman, Thierry Declerck

Overview:
One central aspect of multidimensionality within the context of linguistic data science is being able to represent, interlink, and exchange multimodal linguistic data. Multimodal here refers to the existence of more than one modality, where modality is defined as audio, visual, textual, or other channel.

Activities:
Very soon we established a cooperation with the Ontology Lexica (Ontolex) W3C Community Group for discussing issues related to the representation of multimodal data. We had regular teleconferences with the developers of the Ontolex module for FRequency, Attestations and Corpus data (FrAC), as the representation of multimodal language data was discussed in this group as well.

A first topic discussed in the context of this cooperation concerned the representation of Sign Languages at the lexical level and how to integrate it in the context of OntoLex-Lemon. For this an extensive study of existing approaches for transcribing Sign Language has been proposed. We focused on the HamNoSys notation system (https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/hamnosys-97.html) and its rendering in the SiGML machine readable transformation (Kaur and Kumar, 2016) and (Neves et la., 2020), which can be encoded as the instance of an ontolex:Form class. Parallel to this work, a first version of an ontology for data categories for Sign Languages has been designed. The ontology contains as for today ca 300 classes, instances and properties. It will be published soon, after a round of curation processes.

Another topic concerned the automated addition of pronunciation information to the Open English WordNet data (https://github.com/globalwordnet/english-wordnet). As Wordnets typically do not contain this type of information, but which can be very useful to additionally mark ambiguous Wordnet entries, we attacked this problem in cooperation with the ELEXIS project (https://elex.is/) and extracted pronunciation information form the English edition of Wiktionary, filtered the distinct pronunciations for ambiguous entries (one example being lead (metal element) vs lead (guiding or conduction action). This (disambiguated) pronunciation information has been made available to the Open English WordNet for subsequent integration. See (Declerck and Bajčetić, 2021) for more details.

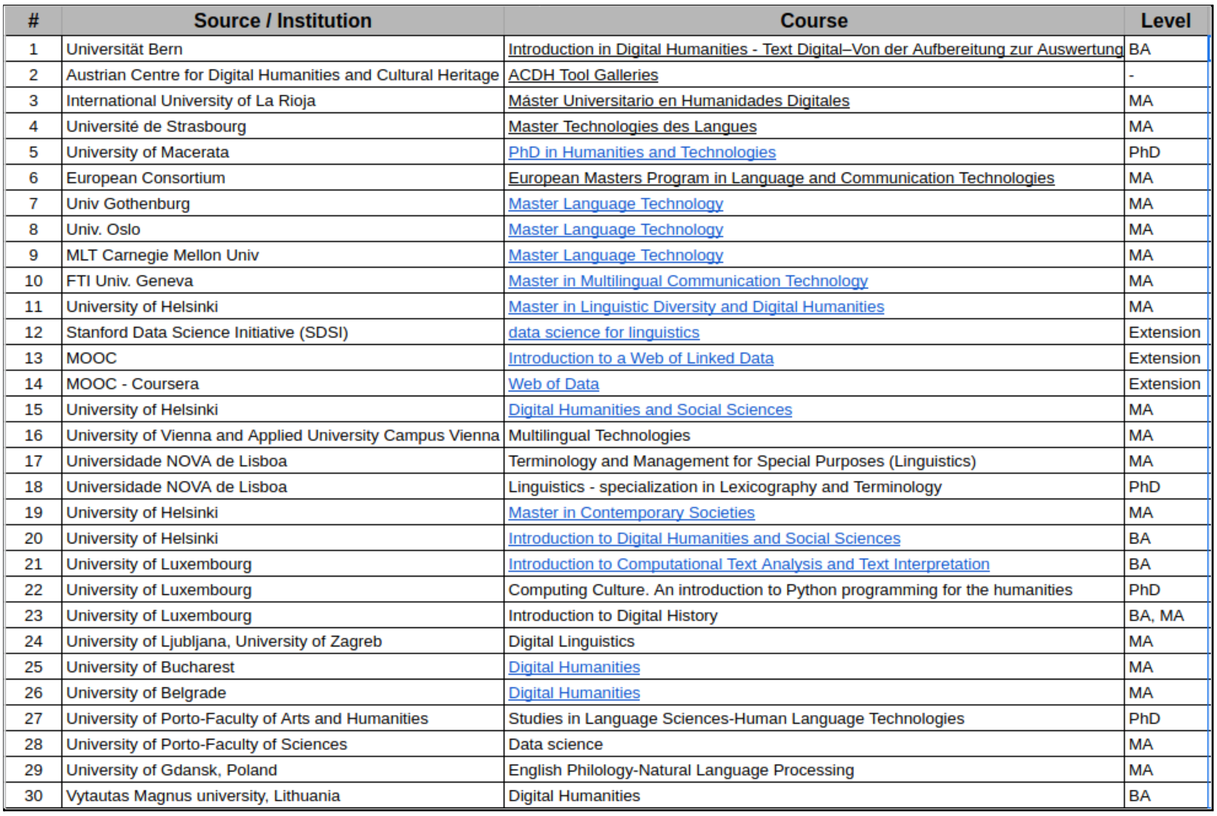
A third topic concerned the processing of data available in writing systems / scripts as most of us currently use, for example in pictographs. Nowadays, such systems are often used in order to facilitate communication between people experiencing problems with ‘our’ writing systems (such as people with intellectual and/or developmental disabilities, migrants wanting to socially integrate and communicate in a language / script they do not yet master, etc) or to communicate with the outside world. Such pictographs can be linked to WordNet, thus enabling translation in other pictograph sets and/or written language. An example we will have a look at are traditional types of Chinese characters, like pictophonetical ones.

3.5. Task 3.5. Education in linguistic data science

Task leaders: Renato Rocha Souza

Overview:
In order to introduce linguistic data science in a cross-disciplinary academic infrastructure, NexusLinguarum is in the process of working out a curriculum for Europe-wide master degree that the participating institutions could adopt to train a new generation of researchers in the area. In spite of being a very broad field, with different vantage points and educational approaches, a curriculum assessment was made in an attempt to map the current academic landscape with the common skills and competencies, and the next steps would involve the actual design of the course as an Erasmus+ initiative.

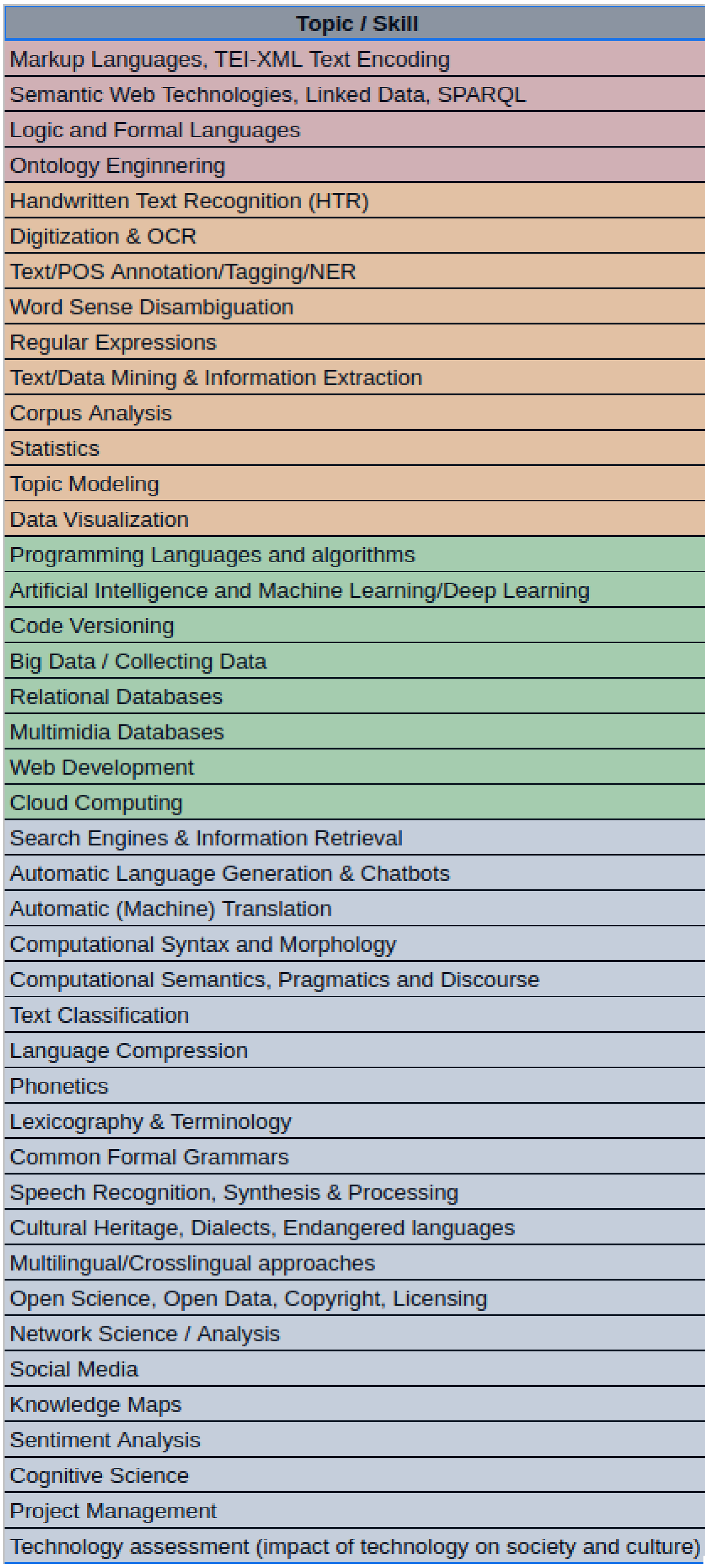
Activities:
The Task 3.5 has begun with an assessment of the existing courses and educational programs related to linguistics and data science, some institutions resulting from this process are exemplified in Table 2. During the Second COST Action meeting in Prague, in 2019, some suggestions were given to constitute an analysis matrix of competencies and skills involved in these programs, into which later suggestions were added.



#	Source / Institution	Course	Level
1	Universität Bern	Introduction in Digital Humanities - Text Digital–Von der Aufbereitung zur Auswertung	BA
2	Austrian Centre for Digital Humanities and Cultural Heritage	ACDH Tool Galleries	-
3	International University of La Rioja	Máster Universitario en Humanidades Digitales	MA
4	Université de Strasbourg	Master Technologies des Langues	MA
5	University of Macerata	PhD in Humanities and Technologies	PhD
6	European Consortium	European Masters Program in Language and Communication Technologies	MA
7	Univ Gothenburg	Master Language Technology	MA
8	Univ. Oslo	Master Language Technology	MA
9	MLT Carnegie Mellon Univ	Master Language Technology	MA
10	FTI Univ. Geneva	Master in Multilingual Communication Technology	MA
11	University of Helsinki	Master in Linguistic Diversity and Digital Humanities	MA
12	Stanford Data Science Initiative (SDSI)	data science for linguistics	Extension
13	MOOC	Introduction to a Web of Linked Data	Extension
14	MOOC - Coursera	Web of Data	Extension
15	University of Helsinki	Digital Humanities and Social Sciences	MA
16	University of Vienna and Applied University Campus Vienna	Multilingual Technologies	MA
17	Universidade NOVA de Lisboa	Terminology and Management for Special Purposes (Linguistics)	MA
18	Universidade NOVA de Lisboa	Linguistics - specialization in Lexicography and Terminology	PhD
19	University of Helsinki	Master in Contemporary Societies	MA
20	University of Helsinki	Introduction to Digital Humanities and Social Sciences	BA
21	University of Luxembourg	Introduction to Computational Text Analysis and Text Interpretation	BA
22	University of Luxembourg	Computing Culture. An introduction to Python programming for the humanities	PhD
23	University of Luxembourg	Introduction to Digital History	BA, MA
24	University of Ljubljana, University of Zagreb	Digital Linguistics	MA
25	University of Bucharest	Digital Humanities	MA
26	University of Belgrade	Digital Humanities	MA
27	University of Porto-Faculty of Arts and Humanities	Studies in Language Sciences-Human Language Technologies	PhD
28	University of Porto-Faculty of Sciences	Data science	MA
29	University of Gdansk, Poland	English Philology-Natural Language Processing	MA
30	Vytautas Magnus university, Lithuania	Digital Humanities	BA

Table 2: Institutions with educational initiatives related to linguistic data science

All the syllabus and detailed program from these courses were analysed in order to extract the main skills and competencies, which were then also listed in an analysis matrix depending on their frequency and quality of occurrence in different programs. The main skills and competencies are depicted in Fig. 6.



Topic / Skill
Markup Languages, TEI-XML Text Encoding
Semantic Web Technologies, Linked Data, SPARQL
Logic and Formal Languages
Ontology Enginnering
Handwritten Text Recognition (HTR)
Digitization & OCR
Text/POS Annotation/Tagging/NER
Word Sense Disambiguation
Regular Expressions
Text/Data Mining & Information Extraction
Corpus Analysis
Statistics
Topic Modeling
Data Visualization
Programming Languages and algorithms
Artificial Intelligence and Machine Learning/Deep Learning
Code Versioning
Big Data / Collecting Data
Relational Databases
Multimidia Databases
Web Development
Cloud Computing
Search Engines & Information Retrieval
Automatic Language Generation & Chatbots
Automatic (Machine) Translation
Computational Syntax and Morphology
Computational Semantics, Pragmatics and Discourse
Text Classification
Language Compression
Phonetics
Lexicography & Terminology
Common Formal Grammars
Speech Recognition, Synthesis & Processing
Cultural Heritage, Dialects, Endangered languages
Multilingual/Crosslingual approaches
Open Science, Open Data, Copyright, Licensing
Network Science / Analysis
Social Media
Knowledge Maps
Sentiment Analysis
Cognitive Science
Project Management
Technology assessment (impact of technology on society and culture)

Figure 6: Main competencies and skills identified

The topics of the list in Fig. 6 were then clustered into thematic groups (e.g. Semantic Web; NLP Pipelines; Coding, Machine Learning & Databases; etc.) and analyzed by frequency of appearance. Some topics would occur in more than 75% of the programs (e.g. "Semantic Web Technologies, Linked Data, SPARQL"; "Text/Data Mining & Information Extraction") while others would appear less frequently (e.g. "Handwritten Text Recognition (HTR)", "Statistics"), according to the program specificities, overarching department/institution and level (e.g. MA, BA, PhD). Some programs have prerequisites, which would demand a deeper analysis on necessary skills. A companion analysis of the software platforms, frameworks and programming languages was also made, yielding the following list of tools: (Oxygen, Tesseract, Transkribus, Git, Protégé, XMLMind, Gephi, R, Sonic, Visualiser, Wordnet, Framenet, Lexonomy, Sklearn - Python package)

Next steps are inviting other COST Action participants to assess the work done so far, for additions, corrections and enlarging the analysed sample. In parallel, the group will study the process for submitting a Erasmus+ proposal, and put forward the initiative. These involve selecting the participating institutions to form a network, fill the documents and templates, select professors, and prepare for the launching of the program.

4. Conclusion

This mid-term activity report details all major activities accomplished in WG3 within the first two years, major collaborations established, major outcomes of these activities, and planned future activities reported in each task description. Apart from all these formal and measurable activities, we would also like to mention the networking and community building effect of this COST Action. In addition to large- and small-scale partnerships on conducting very specific work, jointly preparing publications, or organizing events, we have also established a very friendly and supportive network on the topic of support for linguistic data science, in which expertise is as much exchanged as friendly, amicable conversation (online or now finally also offline again).

References

WG 3 publications

Armaselu, F., Apostol, E. S., Khan, A. F., Liebeskind, C., McGillivray, B., Truică, C. O., & Valūnaitė Oleškevičienė, G. (2021). HISTORIAE, History of Socio-Cultural Transformation as Linguistic Data Science. A Humanities Use Case. In 3rd Conference on Language, Data and Knowledge (LDK 2021). Schloss Dagstuhl-Leibniz-Zentrum für Informatik. https://drops.dagstuhl.de/opus/volltexte/2021/14570/

Declerck, Thierry and Bajčetić, Lenka. (2021). Towards the Addition of Pronunciation Information to Lexical Semantic Resources**.** In *Proceedings of the 11th Global Wordnet Conference, Pages 284-291, Pretoria/Virtual, South Africa, Global Wordnet Association, Global WordNet Association, 1/2021*

Declerck, Thierry. (2020). Towards an Extension of the Linking of the Open Dutch WordNet with Dutch Lexicographic Resources. https://doi.org/10.5281/zenodo.3895201

Declerck, Thierry, McCrae, John, Hartung, Matthias, Gracia, Jorge, Chiarcos, Christian, Montiel, Elena, Cimiano, Philipp, Revenko, Artem, Lee, Deidre, Racioppa, Stefania, Nasir, Jamal, Orlikowski, Matthias, Lanau-Coronas, Marta, Fäth, Christian, Rico, Mariano, Elahi, Mohammad Fazleh, Khvalchik, Maria, Sauri, Roser, Gonzalez, Meritxell, & Katharine Cooney. (2020, May 13). Recent Developments for the Linguistic Linked Open Data Infrastructure. 12th Conference on Language Resources and Evaluation (LREC 2020). https://doi.org/10.5281/zenodo.3934626

Garabik, Radovan (ed.) (2021) Book of Abstract of the Workshop on Deep Learning and Neural Approaches for Linguistic Data, Skopje, North Macedonia, URL: https://www.juls.savba.sk/attachments/workshop_20210930_en/workshop_20210930_book_of_abstracts.pdf.

Gracia, Jorge, Garabík, Radovan & Benko, Vladimír. (2021). NexusLinguarum – European Network for Web-centered Linguistic Data Science. In Slovenská reč, 86(1), pp. 130–132.

Gromann, Dagmar, Lennart Wachowiak, Christian Lang, Barbara Heinisch (2021) Multilingual Extraction of Terminological Concept Systems. In Garabík (ed.) Book of Abstract of the Workshop on Deep Learning and Neural Approaches for Linguistic Data, Skopje, North Macedonia, URL: https://www.juls.savba.sk/attachments/workshop_20210930_en/workshop_20210930_book_of_abstracts.pdf.

Oleškevičienė, Giedrė Valūnaitė, Liebeskind, Chaya, Trajanov, Dimitar, Silvano, Purificação, Chiarcos, Christian & Damova, Mariana (2021). Speaker Attitudes Detection through Discourse Markers Analysis. In Garabík (ed.) Book of Abstract of the Workshop on Deep

Learning and Neural Approaches for Linguistic Data, Skopje, North Macedonia, URL: https://www.juls.savba.sk/attachments/workshop_20210930_en/workshop_20210930_book_of_abstracts.pdf.

Oleškevičienė, Giedrė Valūnaitė & Liebeskind, Chaya (2021). Multiword expressions as discourse markers in Hebrew and Lithuanian. In Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age (pp. 46-56). https://aclanthology.org/2021.motra-1.5/

Oliveira, Hugo Gonçalo (2021). Acquiring Lexico-Semantic Knowledge from a Portuguese Masked Language Model. In Garabík (ed.) Book of Abstract of the Workshop on Deep Learning and Neural Approaches for Linguistic Data, Skopje, North Macedonia, URL: https://www.juls.savba.sk/attachments/workshop_20210930_en/workshop_20210930_book_of_abstracts.pdf.

Păiș, Vasile & Mitrofan, Maria (2021). Towards a named entity recognition system in the Romanian legal domain using a linked open data corpus. In Garabík (ed.) Book of Abstract of the Workshop on Deep Learning and Neural Approaches for Linguistic Data, Skopje, North Macedonia, URL: https://www.juls.savba.sk/attachments/workshop_20210930_en/workshop_20210930_book_of_abstracts.pdf.

Rackevičienė, Sigita, Utka, Andrius, Mockiene, Liudmila, and Rokas, Aivaras (submitted for publication). Methodological Framework for the Development of an English-Lithuanian. In Studies about Languages, Kaunas: KTU.

Rackevičienė, Sigita, Utka, Andrius, Agnė, Bielinskienė, and Rokas, Aivaras (submitted for publication). Distribution of Terms Across Genres in the Lithuanian Cybersecurity Corpus. Respectus Philologicus.

Rokas, Aivaras, Rackevičienė, Sigita, and Utka, Andrius (2020). Automatic extraction of Lithuanian cybersecurity terms using deep learning approaches. In proceedings of the 9th international conference, Baltic HLT, Kaunas, Vytautas Magnus University, Lithuania, 22-23 September 2020, Andrius Utka, Jurgita Vaičenonienė, Jolanta Kovalevskaitė, Danguolė Kalinauskaitė (eds). Amsterdam: IOS Press, 2020.

Susman, Margaux, Leijen, Djuddah, Groom, Nicholas & Johansson, Christer (2021). Investigating Academic Document Structure using Object Detection Methods. In Garabík (ed.) Book of Abstract of the Workshop on Deep Learning and Neural Approaches for Linguistic Data, Skopje, North Macedonia, URL: https://www.juls.savba.sk/attachments/workshop_20210930_en/workshop_20210930_book_of_abstracts.pdf.

Terminology Coordination Unit of the European Parliament (2021). Why is *terminology* your passion? The fifth collection of interviews with prominent terminologists. European Union, 81-86, DOI: 10.2861/665879, URL:

https://www.termcoord.eu/wp-content/uploads/2021/04/L021043-DG-TRAD-ebook-A4-Terminology-5th_PROOF15.pdf

Utka, Andrius, Rackevičienė, Sigita, Mockienė, Liudmila, Rokas, Aivaras, Laurinaitis, Marius, and Bielinskienė Agnė. (accepted for publication) Corpora for Bilingual Terminology Extraction in Cybersecurity Domain. In Proceedings of CLARIN-21 Annual conference.

Other references in this report

Ionov, Maxim, McCrae, John, Chiarcos, Christian, Declerck, Thierry, Bosque-Gil, Julia, & Gracia, Jorge. (2020). Proceedings of the LREC 2020 7th Workshop on Linked Data in Linguistics (1.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3935946

Janev, V., Pujić, D., Jelić, M., & Vidal, M. E. (2020). Survey on Big Data Applications. In *Knowledge Graphs and Big Data Processing* (pp. 149-164). Springer, Cham.

Kaur Khushdeep and Kumar Parteek (2016). HamNoSys to SiGML Conversion System for Sign Language Automation. In Procedia Computer Science, Volume 89, 2016, Pages 794-803, https://www.sciencedirect.com/science/article/pii/S1877050916311280

Neves Carolina, Coheur Luisa & Nicolau Hugo (2020). HamNoSyS2SiGML: Translating HamNoSys Into SiGML. In: Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020). https://aclanthology.org/2020.lrec-1.739/